Chapter 3 Power Analyses for Minimum-Effect Tests

The traditional null hypothesis is that treatments, interventions, etc. have no effect; in Chapters 1 and 2, we use the term "nil hypothesis" to describe this particular version of H_o. The nil hypothesis is so common and so widely used that most researchers assume that the hypothesis that treatments have no effect, or that the correlation between two variables is zero is the null hypothesis. This is wrong. The null hypothesis is simply the specific hypotheses that is being tested (and that might be nullified by the data), and there are an infinite number of null hypotheses researchers might test. One researcher comparing two treatments might test the hypothesis that there is no difference between the mean scores of people who receive different treatments. A different researcher might test the hypothesis that one treatment yields scores that are, on average, five points higher than those obtained using another treatment. Yet another researcher might test the hypothesis that treatments have a very large effect, accounting for at least 25% of the variance in outcomes. These are all null hypotheses. Knowing that there are so many null hypotheses that might be tested, it is useful to understand why one special form – i.e., the nil hypothesis, is the one that actually *is* tested in most statistical analyses.

Nil Hypothesis Testing

There are two advantages to testing the nil hypothesis: (1) it is easy to test this hypothesis- tests of this hypothesis represent the standard method presented in statistics textbooks, data analysis packages, etc., and the derivation of test statistics designed to evaluate the nil hypothesis is often comparatively simple, and (2) if a researcher rejects the hypothesis that treatments have no effect, he or she is left with the alternative that treatments have *some* effect. If H_0 states that nothing happened as a result of treatments, the alternative hypothesis (H_1) is that *something* happened as a result of treatments. In contrast, tests of many specific alternatives to the nil hypothesis can lead to confusing results.

Suppose a researcher tests and rejects that hypothesis that the difference between treatment means is 5.0. The alternative hypothesis (H_1) is that the difference in treatment means is *not* 5.0, and this includes quite a wide range of possibilities. Treatments might lead to differences larger than 5.0. They might have no effect whatsoever. They might have some effect, but lead to differences smaller than 5.0. They might even have the opposite effect than the researchers expected (i.e., a new treatment might lead to worse outcomes than the old one). All a researcher learns by rejecting this null hypothesis is that the difference is not 5.0.

As we noted in Chapters 1 and 2, nil hypothesis testing has been widely criticized (Cohen, 1994; Meehl, 1978; Morrison & Henkel, 1970; Murphy, 1990; Schmidt, 1992, 1996. For discussions of the advantages of this approach, see Chow, 1988; Cortina & Dunlap, 1997; Hagen, 1997). The most general criticism of the traditional approach to null hypothesis testing is that very few researchers actually believe that the nil hypothesis is correct. That is, it is rare to encounter serious treatments or interventions that have no effect whatsoever, which is the traditional null hypothesis. In a later section, we will discuss in detail why the nil hypothesis is almost always false. Here, we will simply note that if the null hypothesis to be tested is known in advance to be false, or is very likely to be false, tests of that hypothesis have very little value (Murphy, 1990).

The second critique of nil hypothesis testing is that the outcomes of tests of the traditional null hypothesis are routinely misinterpreted. As we will show below, the outcomes of standard statistical tests probably say more about the power of your study than about the phenomenon you are studying. If you design a study with sufficient power, you will almost always reject the nil hypothesis. If you design a study with insufficient power, you will usually fail to reject the nil hypothesis. These facts are well understood by most researchers, but it is still common to find that when researchers fail to reject the null hypothesis, they end up drawing conclusions about the treatments or about the relationships being studied. In particular, researchers who fail to reject the null hypothesis all to often conclude that the treatments or interventions being studied did not work, or that the variables being studied are not correlated. Similarly, researchers who *do* reject the null hypothesis often conclude that there are meaningful treatment effects, or that the intervention being studied worked. It should be clear by now that the outcomes of nil hypothesis tests are not driven solely, or even mainly by the substantive phenomenon being studied. The outcomes of nil hypothesis tests usually tell researchers more about the study than they do about the substantive phenomenon being studied.

A third criticism of the traditional null hypothesis is that showing that a result is "significant" at the .05 level does not necessarily imply that it is important or that it is especially likely to be replicated in a future study (Cohen, 1994). This significance test merely shows that the results reported in a study probably would not have been found if the true effect of treatments was zero. Unfortunately, researchers routinely misinterpret the results of significance tests (Cohen, 1994; Cowles, 1989; Greenwald, 1993). This is entirely understandable; most dictionary definitions of "significant" include synonyms such as "important" or "weighty". However, these tests do not directly assess the size or importance of treatment effects, nor do they assess the likelihood that future studies will find similar effects.

Tests of the traditional nil hypothesis are more likely to tell you about the sensitivity of your study than about the phenomenon being studied. With large samples, statistical tests of the traditional nil hypothesis become so sensitive that they can detect the slightest difference between a sample result and the specific value that characterized the null hypothesis, even if this difference is negligibly small. With small samples, on the other hand, it is difficult to establish that *anything* has a statistically significant effect. The best way to get an appreciation of the limitations of traditional null hypothesis tests is to scan the tables in any power analysis book (Cohen, 1988). What you will find is that, regardless of the true strength of the effect, the likelihood of rejecting the traditional null hypothesis is very small when samples are small, and the likelihood of rejecting this hypothesis is extremely high when samples are large. Clearly, there is a need for approaches to significance testing that tell researchers more about the phenomenon being studied than about the size of their samples. Alternatives to traditional nil hypothesis tests are described later in this chapter.

Beginning of Boxed section

Is the Nil Hypothesis Almost Always Wrong ?

There are several reasons to believe that the nil hypothesis is, by definition, almost always wrong (Cohen, 1994; Kirk, 1996). That is, regardless of the treatments interventions, correlations, etc. being studied, it is very unlikely that the true population effect (the difference between two means, the correlation between two variables) is *precisely* zero. First, the hypothesis is in theory usually wrong because is it is a point hypothesis. That is, the hypothesis being tested is that the effect of treatments is exactly zero, even to the millionth decimal place or beyond. The traditional nil hypothesis represents a convenient abstraction, similar to the mythical "friction-less plane" encountered by freshmen in solving physics problems, in which potentially small effects are treated as zero for the sake of simplicity. There are many real-world phenomena that seem to mirror the traditional nil hypothesis, the most obvious being flipping a coin (see Fick, 1995, for examples that seem to show how the traditional null could be correct). However, even in studies that involved repeated flips of a fair coin, the hypothesis is that there is no difference whatsoever in the probability of getting a head or a tail is simply not true. It is impossible to mill a coin that is so precisely balanced that the likelihood of getting heads or tails is exactly equivalent; imbalance at the ten-billionth of the ounce will lead you to favor heads or tails if you flip the coin a sufficient number of times.

Notice that there is nothing special about the hypothesis that the difference between two treatments is zero. All point hypotheses (e.g., that the difference between two treatments is 5.0) suffer from the same problem. In the abstract, they might be true, but in reality, there is no way to ever demonstrate that they are true. Because the nil hypothesis is infinitely precise, none of the real-world phenomena it is designed to test can possibly be assessed at that level of precision. Therefore, if you showed that there was no difference in two treatments, even at the billionth decimal place, the possibility that some difference might emerge with a finer-grained analysis could never be completely dismissed.

A third argument for the conclusion that the nil hypothesis is almost always wrong is best conveyed using a spatial analogy. Suppose you used a standard football field to represent all of the outcomes that could possibly occur when you compare the means of two populations, each of whom received different treatments. One possibility is that there is no difference. Another possibility is that one treatment mean is .01 units higher than the other. Another possibility is that one treatment mean is .02 units higher than the other, and so on. Divide the football field so that H_0 represents the one outcome (no difference) that corresponds with the nil hypothesis and H_1 represents all of the other outcomes. How much space do you think you would devote to H_0 vs. H_1? The most optimistic proponent of nil hypothesis testing would only devote a tiny patch of dirt to H_o. Virtually the entire football field would have to be devoted to H_1 because there is an effectively infinite number of outcomes that could happen in this experiment. H_0 represents one of these outcomes, and H_1 represents all of the rest. In spatial terms, the solution space that represents H_o is essentially infinitely small, whereas the solution space that represents H_1 is infinitely large.

The argument against the traditional null is not only a philosophical one; there are also abundant data to suggest that treatments in the social and behavioral sciences virtually always have at least some effect (Lipsey & Wilson, 1993; Murphy & Myors, 1999). In fact, it may not be possible to devise a real treatment that has no effect whatsoever. To be sure, there is no shortage of crackpot interventions, junk science and treatments that have no meaningful effect (e.g., wearing magnetized bracelets as a way of treating cancer). However, if your goal is to evaluate serious treatments devised by someone who had some sensible reason to believe that they might work, that the likelihood that treatments will have no effect whatsoever is so low that tests of the nil hypothesis may be pointless (Murphy, 1990)

End of Boxed section

Implications of Believing That the Nil Hypothesis is Almost Always Wrong

The fact that the traditional nil hypothesis should almost always be rejected has important implications for thinking about Type I and Type II errors. In fact, we would argue that tests of the traditional nil hypothesis make sense *only* if you believe that the nil hypothesis is often correct, and we know of no researchers who actually believe this. To understand why assumptions about the likelihood that the nil hypothesis is true are so important, consider Figure 3-1, which illustrates two ways of making errors in a nil hypothesis test. First, it is possible that H_o is really true, and that researchers will wrongly conclude that treatments do have some effect – i.e., a Type I error. There is a large and robust literature detailing methods for controlling Type I errors (e.g., Wilkinson et al., 1999; Zwick & Marascuilo,1984). As Figure 3-1 makes clear, the likelihood of making a Type I error depends first and foremost on whether or not the nil hypothesis that treatments have no effect is actually true (Murphy, 1990). If you believe that the nil hypothesis is never true, it follows that Type I errors are

Figure 3-1
Errors Statistical Tests

		What is True in the Population ?	
		Treatments Have No Effect	Treatments Have Some Effect
Conclusions Reached in A Study	No Effect		Type II Error
	Treatment Effect	Type I Error	

impossible – these errors can occur if and only if H_o is true. If you believe that the nil hypothesis is almost always wrong, it follows that Type I errors are at best very rare. Unless you believe that he nil hypothesis is often right, it is unlikely that you will ever have much reason to be concerned about Type I errors.

Beginning of Boxed section

Polar Bear Traps: Why Type I Error Control is a Bad Investment

A simple analogy helps to drive home the importance of the low likelihood that the nil hypothesis is true. Suppose you are a homeowner and a salesman comes to the door selling Polar Bear Traps. He makes a compelling case that a polar bear attack would be very unpleasant and that these are good traps. Would you buy traps? We would not. The likelihood that a polar bear will attack is so small that the traps strike us as a waste of money.

You should think of any and all procedures designed to control or reduce Type I errors as Polar Bear Traps. They give you a means of controlling or limiting something that you know is highly unlikely in the first place. More important, these traps cost something. Virtually anything you do to control Type I errors will increase the likelihood of Type II errors. For example, the common rationale for using a stringent alpha level, such as α = .01 rather than a less demanding criterion (e.g., α = .05) in defining statistical significance is that the stricter alpha will lead to fewer Type I errors. If you think that Type I errors are virtually impossible in the first place, you are unlikely to believe that the loss of power implied by the more stringent alpha is worthwhile. In general, steps you take to control Type I errors usually lead to reduced levels of power. If you think about Type I error control as a Polar Bear Trap, you are unlikely to see it as a wise investment.

Beliefs and arguments about the truth or falsity of the nil hypothesis are exactly that. In principle, we do not think you could conclusively demonstrate that the nil hypothesis has either true or false in a particular setting, regardless of the data you had at hand. As a result, different scientists may come to different conclusions about whether the nil hypothesis is always or even almost always false. However, it is important to bear in mind that the whole logic of traditional nil hypothesis testing only holds together if you believe that there is a realistic possibility that the nil hypothesis is actually true in your study, and it is hard to find researchers who seriously believe this proposition.

End of Boxed section

Throughout this chapter, we will show how the conclusion that the traditional nil hypothesis is rarely true undermines many familiar methods of statistical analysis. If a you believe that the nil hypothesis is never true, it follows that tests of that hypothesis *never* tell you anything about the substantive questions your study is trying to answer (although they may tell you some things about the design of the study). If you accept our argument that the null hypothesis is almost always wrong in serious research studies, it follows that tests of this hypothesis have very little value, and that alternatives are needed that will solve some of the problems inherent in this method of hypothesis testing. If you truly believe that the nil hypothesis is so often true that tests of this hypothesis have any real probative value, you are in a very small minority. Most researchers who routinely use nil hypothesis tests simply do not understand that they are using procedures that make sense only in the unlikely case where H_o is actually likely to be true.

Beginning of Boxed section

<u>The Nil may not be True, but it is Often Fairly Accurate</u>

The conclusion that the nil hypothesis is almost always wrong is not the same as the conclusion that most treatments, interventions, etc. actually work. On the contrary, it is reasonable to believe that many treatments and interventions have very small effects, and that the statement that they had no effect at all is often fairly close to the truth. As we noted in Chapters 1 and 2, most studies you read in the literature are likely to deal with treatments, interventions, etc. that have small (but perhaps important) effects. If the effects of treatments were large and obvious, there would not be much demand for further studies testing the nil hypothesis. Many of the most important and interesting studies you are likely to read will deal with new treatments, novel interventions, and innovative approaches, and many of these just won't work very well. As a point hypothesis, the nil hypothesis is almost always wrong, but as a general description of what researchers expect might happen, it is often a good approximation. What is needed is an approach that captures the worrisome possibility that many treatments, interventions, etc. might have very small effects, without all of the problems that accompany tests of point hypotheses. A method that overcomes many of the limitations of traditional nil hypothesis tests is presented in the section that follows.

End of Boxed section

Minimum-Effect tests as Alternatives to Traditional Null Hypothesis Tests

The criticisms of nil hypothesis tests outlined above have led some critics to call for abandoning null hypothesis testing altogether (e.g., Schmidt, 1992, 1996). Rather than take this drastic step, we think it is better to reform the process. In our view, the most serious problem with most null hypotheses tests is that the specific hypothesis being tested, i.e. - that treatments have no effect whatsoever, is neither credible nor informative (Murphy, 1990). There are several alternatives to testing the "nil hypothesis", and all of these are a marked improvement over the standard procedure of testing the hypothesis that the effect of treatments is precisely zero.

Serlin and Lapsley (1993) show how researchers can test the hypothesis that the effect of treatments falls within or outside of some range of values that is "good enough" to establish that one treatment is meaningfully better than the other. Rouanet (1996) shows how Bayesian methods can be used to assert the importance or negligibility of treatment effects. Both of these methods allow researchers to directly test credible and meaningful hypotheses.

The method described in this book involves testing “minimum effect” hypotheses (Murphy & Myors, 1999). The traditional nil hypothesis involves a choice between:

H_0 – treatments have no effect

vs.

H_1 – treatments have some effect

In contrast, minimum effect hypotheses involve a choice between:

H_0 – treatments have an effect that is so small that it can be described as negligible

vs.

H_1 – treatments have an effect that is so large that they should be described as meaningful or important

Minimum-effect tests require researchers to make a decision about what sort of treatment effect is so small that it should be labeled as negligible. This decision may be a difficult one, but if a reasonable standard for defining effects that are negligibly small can be defined, it is quite easy to develop the appropriate tests, using the same noncentral-*F*-based model we have used in presenting power analysis.

The central assumption of minimum effect tests is that there is some range of possible ES values that are so small that they might as well be zero. For example, if there is a substantive reason to believe that a treatment effect that accounts for less than 1% of the variance is simply too small to care about, it does not matter whether the true treatment effect is *PV* = 0.0, *PV* = 0.001, *PV* = 0.002, etc. Anything that falls below *PV* = 0.01 will be regarded as too small to be meaningful or important, and the null hypothesis in this minimum-effect test is that the population treatment falls somewhere within this range. A researcher who can reject this null hypothesis is left with the alternative hypothesis that effects are large enough to be important.

Minimum-effect tests have many advantages over traditional nil hypothesis tests. First, both H_0 and H_1 are substantively meaningful and intrinsically interesting to most researchers. Second, the minimum-effect null hypothesis, that treatments have effects so small that they are negligible, is one that might actually be true. Third, these tests can be developed in ways that allow researchers to apply virtually all of the procedures, standards, and metrics they have learned in the context of nil hypothesis testing to tests that the effects of treatments are either negligibly small or large enough to be of interest.

Minimum-effect tests are meaningful. There is much to be learned by conducting tests the minimum-effect null hypothesis that effects of treatments are negligibly small (e.g., they account for 1% or less of the variance in outcomes). In contrast to tests of the traditional null, tests of this sort are far from trivial. First, when conducting minimum-effects tests, researchers do not know in advance whether H_0 is right or wrong. Second, these tests involve questions that are of real substantive interest. Because these include some range of values rather than a single exact point under the "null umbrella", the results of these tests are not a foregone conclusion. Although it might be impossible to devise a treatment, intervention, etc. that had *no* effect whatsoever, there are any number of treatments whose effects fall somewhere between zero and whatever point you choose to designate as a negligible effect. The possibility that your treatments will have a negligibly small effect is both real and meaningful (whereas the possibility that they will have no effect whatsoever is not), and researchers can learn something important about their treatments by testing this hypothesis.

The minimum-effect might actually be true. In an earlier section of this chapter, we noted that when testing the traditional null hypothesis, a researcher can guarantee the outcome if the sample is sufficiently large. That is, with a sufficiently sensitive study, the statistical power of tests of the traditional null hypothesis will reach 1.00. The reason it is possible to reach power levels of 1.00, regardless of the substantive question being examined in a study, is that traditional models of power analysis are based on the entirely realistic assumption that H_o is virtually never true.

As we noted above, the assumption that H_o is virtually never true also means that Type I errors are virtually impossible and that significance tests will rarely tell you anything meaningful about the substantive questions being studied. In contrast to the nil hypothesis, the minimum-effect null hypothesis is often a realistic possibility. That is, there certainly are treatments, interventions, etc. that turn out to have truly small effects. Because there is a real possibility that the effect of treatments will turn out to be negligible, H_0 might really be true, and Type I errors might really occur.

Because there is a very real possibility that the effects of treatments will turn out to be negligibly small, the upper bound of the power of minimum-effects tests will generally be lower than 1.00. No matter how sensitive the study, researchers can never be certain in advance of the result of a minimum-effect test. We regard this as a good thing. One major criticism of tests of the traditional nil hypothesis is that they are literally pointless (Murphy, 1990). Because researchers know in advance that the traditional H_0 is almost certainly wrong, they are unlikely to learn anything new about H_0, regardless of the outcome of the significance test. Minimum-effect tests, on the other hand, can be informative, particularly if these tests are conducted with acceptable levels of statistical power. If a researcher designs a study that has a great deal of power and still fails to reject the hypothesis that a the effects of treatments are negligible, this failure to reject the null hypothesis can be interpreted as strong evidence that the treatment effect truly *is* negligible.

Minimum-effect tests produce meaningful information. Tests of the nil hypothesis do not necessarily tell researchers anything useful about the phenomenon being studied. Consider a study comparing two different methods of mathematics instruction. Students are randomly assigned to one of these two methods, and at the end of the year, they all take the same final examination. A study reports a significant difference ($\alpha < .05$) between the mean test scores in these two treatments. What have we really learned?

Without additional information (e.g., sample size), the result presented above does not necessarily tell you anything meaningful about the two treatments. For example, it is possible that there is a miniscule difference in the mean scores of students who receive these two treatments, and that the researcher used a sufficiently large sample to detect that difference. It is possible that the difference is large, but the significance test does not, by itself, tell you that.

Suppose another researcher fails to reject the null hypothesis. Does this mean that there is literally no difference between the two treatments? Again, in the absence of additional information, it is impossible to say what this study tells us about the treatments. If the sample was very small, the researcher will not reject H_o - even if one method is truly much better than another.

Suppose another researcher examined these same two treatments, and determines that treatment effects that account for 1% or less of the variance in some important outcome can safely be labeled as negligible. If this researcher rejects the null hypothesis that treatment effects are negligible (i.e., the population effect falls somewhere between *PV* =0.0 and *PV* =.01, this researcher has learned something meaningful about the treatments. In particular, this study has presented evidence that there is a meaningful difference between the two methods of mathematics instruction.

Beginning of Boxed Section

How do you Know the Effect Size?

You have probably noticed that there is some circularity in many discussions of power analysis. To do a good job in power analysis, you need to know the effect size in the population (e.g., how much of the variance in job performance can be explained by individuals differences in general cognitive ability). However, if you *knew* this effect size, you probably would not have needed to do your study in the first place. More to the point, if you are really quite certain about the value of an effect size in the population, it is unlikely that whatever study you do will lead you to change your mind about the size of the effect you are studying (Newman, Jacobs & Bartram, 2007).

In Chapter 1, we observed that in power analysis, precision is not always necessary or even useful. This is especially the case when there is not good basis for estimating the effect size in the population (a recent and comprehensive meta-analysis is an example of such a basis). What is a researcher to do?

Our advice is to be conservative in estimating population effect size values. If you design a study that has sufficient power to test a hypothesis when the population *PV* = .02, that study will have even more power for than your calculations suggest if the population ES is larger than *PV* = .02. On the other hand, over-estimating the size of the effect in a population could lead you to design and execute studies that have less power than you think they have. Starting with a conservative estimate of the ES covers just about all of the bases.

End of Boxed Section

Testing the Hypothesis that Treatment Effects are Negligible

The best way to describe the process of testing a minimum-effect hypothesis is to compare it to the process used in testing the traditional nil hypothesis. The significance of the *F* statistic is usually assessed by comparing the value of the *F* obtained in a study to the value listed in a standard *F* table. The tabled values presented in virtually every statistics text correspond to specific percentiles in the central *F* distribution. For example, if df_{hyp} =2 and df_{err} =100, the tabled values of the *F* statistic that is used in testing the nil hypothesis are 3.09 and 4.82, for α = .05 and α = .01, respectively. In other words, if the nil hypothesis is true and there are 2 and 100 degrees of freedom, you should expect to find *F* values of 3.09 or lower 95% of the time, and values of 4.82 or lower 99% of the time. If the *F* in a particular study is larger than these values, the researcher can reject the null hypothesis and conclude that treatments probably do have some effect.

Tests of minimum-effect hypotheses proceed in exactly the same way, the only difference being that they use a different set of tabled *F* values (Murphy & Myors, 1999). The *F* tables found in the back of most statistics texts are based on the central *F* distribution, or the distribution of the *F* statistic that would be expected if the traditional nil hypothesis were true. Tests of minimum-effect hypotheses are based on a noncentral *F* distribution rather than the central *F* that is used in testing the nil hypothesis.

For example, suppose a researcher decides that treatments that account for 1% or less of the variance in outcomes have a "negligible" effect. It is then possible to estimate a noncentrality parameter (based on *PV* = .01), and to estimate the corresponding noncentral *F* distribution for testing the hypothesis that treatment effects are at best negligible. If *PV* = .01, df_{hyp} =2, and df_{err} =100, 95% of the values in this noncentral *F* distribution will fall at or below 4.49 and 99% of the values in this distribution will fall at or below 6.76 (as we note below, *F* values for testing minimum-effect hypotheses are listed in Appendix B). In other words, if the observed *F* in this study is greater than 4.49, the researcher could be confident (α = .05) in rejecting the hypothesis that treatments accounted for 1% or less of the variance. Later in this chapter, we will discuss standards that might be used in designating effects as "negligible".

You might notice that minimum-effect hypotheses involve specifying a range of values as "negligible". In the example above, effects that account for 1% or less of the variance in the population were designated as negligible effects, and if a researcher can reject the hypothesis that the effects are negligibly small, he or she is left with the alternative hypothesis that they are *not* negligibly small - i.e., that treatment effects are large enough to care about.

You might wonder how a single critical *F* value allows you to test for a whole range of null possibilities. One characteristic of the *F* statistic is that it ranges from zero to infinity, with larger *F* values indicating larger effects. Therefore, if you can be 95% confident that the observed *F* in your study is larger than the *F* you would have obtained if treatments accounted for 1% of the variance, you can also be at least 95% confident that the observed *F* would be larger than that which would have been obtained for any *PV* value between .00 and .01. If the observed *F* is larger than the *F* values expected 95% of the time when *PV* = .01, it must also be larger than 95% of the values expected for any smaller *PV* value.

An example. Suppose 125 subjects are randomly assigned to one of five treatments. You find *F* (4,120) = 2.50, and in this sample, treatments account for 7.6% of the variance in the dependent variable. This *F* is large enough to allow you to reject the traditional null hypothesis (α = .05; the critical value of *F* for testing this nil hypothesis is *F* = 2.45). This significant *F* allows the researcher to reject the nil hypothesis, but since the nil hypothesis is almost always wrong, rejecting it in this study does not really tell the researcher much about the treatments.

Suppose also that treatments of this sort that account for less than 1% of the variance in the population have effects that can sensible be labeled as "negligible". To test the hypothesis that the effects observed in this study came from a population in which the true effect of treatments is negligibly small, all the researcher needs to do is to consult the noncentral F distribution with df_{hyp} =4, df_{err} =120, and λ=1.21 (i.e., a good estimate of the noncentrality parameter λ is given by [120 * .01]/[1 - .01]). In this noncentral *F* distribution, 95% of the values in this distribution are 3.13 or lower. The obtained *F* was 2.50, which is smaller than this critical value. This means that the researcher cannot reject this minimum-effect null hypothesis. That is, although the researcher can reject the hypothesis that treatments have no effect whatsoever (i.e., the traditional null), the researcher cannot reject the hypothesis that the effects of treatments are negligibly small (i.e., a minimum-effect hypothesis that treatments account for less than 1% of the variance in outcomes).

Defining a minimum effect. The main advantage of the traditional nil hypothesis is that it is simple and objective. If a researcher rejects the hypothesis that treatments have no effect, he or she is left with the alternative that they have some effect. On the other hand, testing minimum-effect hypotheses requires value judgments, and requires that some consensus be reached in a particular field of inquiry. For example, the definition of a "negligible" effect might reasonably vary across areas, and there may be no set convention for defining which effects are so small that they can be effectively ignored and which cannot. An effect that looks trivially small in one discipline might look reasonably large in another. However, it is possible to offer some broad principles for determining when effects are likely to be judged "negligible".

First, the importance of an effect should depend substantially on the particular dependent variables involved. For example, in medical research it is common for relatively small effects (in terms of the percentage of variance explained) to be viewed as meaningful and important (Rosenthal, 1993). One reason is that the dependent variables in these studies often include quality of life, and even survival. A small percentage of variance might translate into many lives saved.

Second, decisions about what effects should be labeled as "negligible" might depend on the relative likelihood and relative seriousness of Type I vs. Type II errors in a particular area. As we will note in a section that follows, the power of statistical tests in the general linear model decreases as the definition of a "negligible effect" expands. In any particular study, power is higher for testing the traditional nil hypothesis that treatments have no effect than for testing the hypothesis that they account for 1% or less of the variance in outcomes, and higher for tests of the hypothesis that treatments account for 1% or less of the variance than for tests of hypothesis that treatments account for 5% or less of the variance in outcomes. If Type II errors are seen as particularly serious in a particular area of research, it might make sense to choose a very small figure as the definition of a "negligible" effect.

On the other hand, there are many areas of inquiry in which numerous well-validated treatments are already available (See Lipsey & Wilson, 1993, for a review of numerous meta-analyses of treatment effects in the behavioral sciences. The Cochrane Collaboration has conducted hundreds of review in a range of health-related areas. See http://www.cochrane.org), and in these areas, it might make sense to "set a higher bar" by testing a more demanding hypothesis. For example, in the area of cognitive ability testing (where the criterion is some measure of performance on the job or in the classroom), it is common to find that tests account for 20-25% of the variance in the criterion (Hunter & Hunter, 1984; Hunter & Hirsch, 1987). Tests of the traditional null hypothesis (i.e., that tests have no relationship whatsoever to these criteria) are relatively easy to reject; if $r^2 = .25$, a study with $N = 25$ will have power of .80 for rejecting the traditional null hypothesis (Cohen, 1988). Similarly, the hypothesis that tests account for 5% or less of the variance in these criteria is easy to reject; if $r^2 = .25$, a study with $N = 55$ will have power of .80 for rejecting this minimum-effect hypothesis (Murphy & Myors, 1999). In this context, it might make sense to define a "negligible" relationship as one in which tests accounted for 10% or less of the variance in these criteria.

Utility analysis has been used to help determine whether particular treatment have effects that are large enough to warrant attention (Landy, Farr & Jacobs, 1982; Schmidt, Hunter, McKenzie & Muldrow, 1979; Schmidt, Mack & Hunter, 1984). Utility equations suggest another important parameter that is likely to affect the decision of what represents a negligible vs. a meaningful effect - i.e., the standard deviation of the dependent variable, or SD_y. When there is substantial and meaningful variance in the outcome variable of interest, a treatment that accounts for a relatively small *percentage* of variance might nevertheless lead to practical benefits that far exceed the costs of the treatment.

For example, suppose a training program costs $1,000 per person to administer, and that it is proposed as a method of improving performance in a setting where the current SD_y (i.e., the standard deviation of performance) is $10,000. Utility theory can be used to determine whether to compare the costs with the expected benefits of this method of training. Depending on the effectiveness of the training program, researchers might conclude that benefits exceed costs (if the training is highly effective) or that costs exceed benefits (if the training is not very effective). Utility theory equations can also be used to determine the level of effectiveness the training program needs to meet so that benefits will at least equal costs, and it can be argued that this effectiveness level represents a very good definition of a minimum effect. That is, training programs that lead to more costs than benefits are not likely to be seen as effective, whereas training programs that lead to benefits that exceed their costs will be seen as effective. This break-even point represents a sensible definition of a minimally effective program.

Landy, Farr and Jacobs (1982) discuss the application of utility theory in evaluating performance improvement programs. They note that the overall benefit of such interventions can be estimated using the equation:

$$\Delta U = (r_{xy} * SD_y) - C \qquad [11]$$

Where:

ΔU = the projected gain in productivity associated with training

r_{xy} = the correlation between training and performance

C = the cost of the training program

In this example, SD_y = $10,000 and C = $1,000, so Equation 11 can be restated as:

$\Delta U = (r_{xy} * \$10{,}000) - \$1{,}000$ [12]

A minimally effective training program will produce benefits that are equal to costs. The benefits of this training program depend on its effectiveness - i.e., the benefits are estimated by ($r_{xy} *$ \$10,000). Therefore, in defining the minimum level of training effectiveness that will allow benefits to offset costs, all we need to do is to determine the value of r_{xy} at which predicted benefits equal costs.

Benefits equal costs when ($r_{xy} * SD_y$) equals C. Rearranging the terms in Equation 11, this break-even point is defined as:

$(r_{xy} * SD_y) = C$ when $r_{xy} = C / SD_y$ [13]

That is, benefits equal costs when the effectiveness of training (r_{xy}) is equal to C / SD_y. In our example, benefits equal costs when r_{xy} is equal to .10 (i.e., \$1,000 / \$10,000). If this value of r_{xy} is squared, the point at which the benefits of training at least equal costs is when *PV* = .01. In other words, if training accounts for at least one percent of the variance in performance, benefits will at least offset costs, and this strikes us as a sensible definition of a minimum effect.

In many of the examples presented in this chapter and chapters that follow, we will use conventions similar to those described in Cohen (1988), describing treatments that have less than 1% of the variance as having small effects, and those that account for less than 5% of the variance in outcomes as having small to medium effects. Many of the tables presented in this book are arranged around these particular conventions. However, it is critical to note that the decision of what represents a "negligible" effect is one that is likely to vary across research areas, and that there will be many cases in which these particular conventions do not apply. Appendix B presents the information needed to determine critical *F* values for minimum-effect tests that employ some other operational definition of "negligible", and we urge researchers to carefully consider their reasons for choosing any particular value as a definition of the minimum effect of interest.

Power of minimum-effect tests. As this example suggests, researchers should expect less power when testing the hypothesis that effects exceeds some minimum value than when testing the hypothesis that the effect exactly zero. Switching from a central to a noncentral distribution as the basis of your null hypothesis necessarily increases the *F* value needed to reach significance, thereby reducing power.

The traditional hypothesis that treatments have no effect whatsoever is almost always wrong (Murphy, 1990), and is therefore relatively easy to reject. If the sample is large enough, researchers will *always* reject the hypothesis that treatments have no effect, even if the true effect of treatments is extremely small. Tests of the hypothesis that treatment effects exceed the standard that is used to define "negligible" are more demanding than tests of the traditional nil hypothesis, in part because there is always a chance that the treatment effects *are* negligible. Therefore, there is no guarantee that a researcher will reject a minimum-effect hypothesis, no matter how sensitive the study. However, as we note in several of the chapters that follow, the lower power of minimum-effect tests is easily offset by the fact that these tests tell researchers something meaningful, regardless how large the sample size or how sensitive the study.

Using the One-Stop Tables to Assess Power to Test Minimum-Effect Hypotheses

In Chapter 2, we described the use of the "One Stop *F* Table", presented in Appendix B and the “One-Stop *PV* Table”, presented in Appendix C, in performing significance tests and power analyses for nil hypothesis tests. These same tables also include all of the information you need to conduct significance tests and power analyses for tests of minimum-effect hypotheses, where negligible effects are defined as those that account for 1% or less of the variance, or as those that account for 5% or less of the variance in outcomes.

Testing minimum-effect hypotheses (PV=.01). Rather than testing the hypothesis that treatments have no effect whatsoever, researchers might want to test the hypothesis that treatment effects are so small that they account for less than 1% of the variance in outcomes. If this *PV* value represents a sensible definition of a "negligible" effect in a particular area of research, and researchers test and reject this hypothesis, they can be confident that effects are *not* negligibly small.

The fifth and sixth values in each cell of the One-Stop *F* table are the critical *F* values needed to achieve significance (at the .05 and .01 levels respectively) when testing the hypothesis that treatments account for 1% or less of the variance in outcomes. With df_{hyp} =3 and df_{err} =50, an *F* of 3.24 or larger is needed to reject (with α = .05) the hypothesis that the treatment effect is negligibly small (defined as accounting for 1% or less of the variance in outcomes).

The seventh and eighth values in each cell are *F* - equivalents of the effect size values needed to obtain particular levels of power (given an α level of .05 and the specified df_{hyp} and df_{err}) for testing this minimum-effect hypothesis. The values in the table are 2.46 and 4.48, for power levels of .50 and .80, respectively. [1] This translates into *PV* values of .13 and .21, respectively. That is, if treatments accounted for 13% of the variance in the population, a study with that uses minimum effect tests with df_{hyp} =3, df_{err} =50 and α=.05 will have power of about .50 for detecting that effect. If treatments account for 21% of the variance in the population, the power of this study will be approximately .80.

Testing minimum-effect hypotheses (PV=.05). There are many treatments that routinely demonstrate moderate to large effects. For example, well-developed cognitive ability tests allow researchers to predict performance in school and in many jobs with a relatively high degree of success (correlations in the .30 - .50 range are common). Rather than testing the hypothesis that tests have no relationship whatsoever with these criteria (i.e., the traditional nil), or even that treatments account for 1% or less of the variance in outcomes, it might make sense to test a more challenging hypothesis - i.e., that the effect of this particular treatment is at least small to moderate in size. For reasons that are explained in the section that follows, there are many contexts in which it is useful to test the hypothesis that treatments explain 5% or less of the variance in the population. If a researcher can test and reject the hypothesis that effects explain 5% or less of the variance in outcomes, he or she is left with the alternative that treatments explain more than 5% of the variance. In most contexts, effects this large are likely to be treated as meaningful, even if smaller effects (e.g., those accounting for 1% of the variance) are not.

The ninth and tenth values in each cell of the One-Stop *F* table represent the critical *F* values needed to achieve significance (at the .05 and .01 levels respectively) when testing the hypothesis that treatments account for 5% or less of the variance in outcomes. With df_{hyp} =3 and df_{err} =50, an F of 4.84 or larger is needed to reject (with α = .05) the hypothesis that treatments account for 5% or less of the variance in the population.

The eleventh and twelfth values in each cell are *F* - equivalents of the effect size values needed to obtain particular levels of power (given an α level of .05 and the specified df_{hyp} and df_{err}) for tests of this minimum-effect hypothesis. The values in the table are 4.08 and 6.55, for power levels of .50 and .80, respectively. This translates into *PV* values of .19 of and .28, respectively. In other words, in order to have power of .50 for testing the hypothesis that treatments account for 5% or less of the variance in outcomes in a study with df_{hyp} =3 and df_{err} =50, the true population effect size will have to be fairly large (*PV* = .19). In order to achieve power of .80, the population effect will have to be quite large (*PV* = .28).

Beginning of Boxed Section

A Worked Example of Minimum-Effect Testing

A researcher randomly assigns 125 participants to one of five treatments, and reports *F* (4,120) = 3.50, and that treatments account for 10% of the variance (if you apply formula 6 in Chapter 2, you can confirm this transformation of *F* to *PV* - i.e, *PV* = (df_{hyp} * *F*) / [(df_{hyp} * *F*) + df_{err}]). What conclusions can you draw from this study?

First, it is possible to reject the nil hypothesis, at both the .05 and .01 levels; the critical values shown in Appendix B for df_{hyp} =4 and df_{err} =120 are 2.45 and 3.48, respectively. If the observed value of *PV* (i.e., *PV* = .10), is taken as a reasonable estimate of the population *PV*, the study had more than adequate power for these nil hypothesis tests. As Appendix C shows, a population *PV* of .09 would be large enough to provide power of .80 for tests of the nil hypothesis when df_{hyp} =4 and df_{err} =120.

Second, it is possible to reject the null hypothesis (α = .05) that treatments account for 1% or less of the variance. Appendix B shows that an *F* value of 3.13 or greater would be needed to reject this null hypothesis. If you set a more stringent alpha level (e.g., α = .01), the observed *F* will be smaller than the critical value of *F* (i.e., *F* = 4.40). In other words, the researcher will not be able to reject, with a 99% level of confidence, the hypothesis that treatments account for 1% or less of the variance.

Appendix F shows that population effect sizes of *PV* = .07 and *PV* =.11 will be needed to achieve power levels of .50 and .80, respectively, in tests of this minimum-effects hypothesis. The observed *PV* falls between these two values, and if this observed effect is used as an estimate of the population effect, this suggests that power is slightly lower than .80.

Finally, the results of this study would not allow researchers to reject (α = .05) the hypothesis that treatments account for 5% or less of the variance. The critical value of *F* for this null hypothesis test is 5.45. In order to have power of .80 for testing this hypothesis with df_{hyp} =4 and df_{err} =120, the population treatment effect would have to be quite large (i.e., *PV* = .18).

End of Boxed section

You might note that our One-Stop *F* Table does not contain a set of rows corresponding to relatively large effects (i.e., effects accounting for more than 5% of the variance in outcomes). As we noted in Chapter 1, when the effect of treatments is known or thought to be large, there is often no point to conducting the research. Large effects are usually so obvious that a study confirming their existence is unlikely to make much of a contribution. More to the point, when the effects under consideration are large, statistical power is unlikely to be a problem unless samples are extremely small. When samples are this small, researchers have problems that are much more severe than statistical power (e.g., lack of generalizability), and power analyses for large effects strike us as very limited in value.

Using the One-Stop F Calculator for Minimum-Effect Tests

Both the One-Stop *F* Table and the One-Stop *PV* Table (Appendices B and C) are designed to support both nil hypothesis tests and minimum-effect tests. In particular, these tables provide the information needed to carry out significance tests and power analyses when the hypothesis being tested is that treatments account for 1% or less of the variance, or that treatments account for 5% or less of the variance in outcomes. The One-Stop *F* calculator is also designed to support both nil hypothesis and minimum-effect hypothesis tests. Unlike the tables presented in Appendices B and C, this calculator allows researchers to conduct significance tests and power analyses for all possible minimum-effect tests.

Suppose the working definition of a negligible effect in a particular area of research was an effect that accounted for less than 2.5% of the variance in outcomes. There is nothing sacred about values of 1% or 5% of the variance as definitions of "negligible", and any other value might be chosen. It would be possible to conduct minimum-effect tests for this particular definition of "negligible" by calculating the noncentral *F* distribution that corresponds to your test. A simpler method is to use the One-Stop *F* Calculator, which is designed to make the process of minimum-effects testing every bit as easy as testing the traditional nil hypothesis.

Suppose you randomly assign 102 participants to either a training program that requires active learning or a training program based on traditional lectures. You find a significant difference between the mean test scores of these two groups, *t* =3.66, *PV* =

.118. The first step in applying the models described in this chapter is to transform this *t* value into an *F*. Since $F = t^2$, this translates into *F* = 13.39, with $df_{hyp} = 1$ and $df_{err} = 100$.

To test for statistical significance ($\alpha = .05$), using a minimum-effects test in which effects accounting for 2.5% of the variance or less are treated as negligible, choose the "Significance Testing" option of the One-Stop *F* Calculator and enter values of .025, .05, 1, and 100 in the Effect Size, Alpha, df_{hyp} and df_{err} boxes. The Calculator indicates that the critical value of *F* for testing this hypothesis is 10.87 (the corresponding *PV* is .10). The observed *F* is larger than the critical value, so you can be confident that the in the population, treatments account for more than 2.5% of the variance.

The "Power Analysis" option of this calculator suggests that the power of this study is lower than .80. If you enter values of .025, .80, .05, 1 and 100 in the Effect Size, Power, Alpha, df_{hyp} and df_{err} boxes, you will find that a population *PV* of .135 would be needed to yield power of .80. The observed *PV* is .116, substantially smaller than the value needed to achieve power of .80.

If the researcher wanted to achieve power of .80, he or she would need a lager sample. The "Sample Size (df Error) Determination" option of the calculator allows the researcher to determine sample size requirements. Enter .025, .80, .05, 1. And .118 in the Effect Size, Power, Alpha, df_{hyp} and *PV* boxes, and you will find that df_{err} must be at least 123 ($df_{err} = N$-2, so *N* must be 125) to achieve power of .80 for testing this null hypothesis.

Beginning of Boxed Section

Does Another Perspective on Type I and Type II Errors Solve the Problem that the Traditional Nil Hypothesis is False in Most Cases?

In traditional null hypothesis testing, a great deal of attention has been paid to the likelihood of making Type I (a) vs. Type II (b) errors. You make a Type I error when you reject the null hypothesis when it is in fact true, and you make a Type II error when you fail to reject the null hypothesis when you should (because it is not true).

We have argued on a variety of logical bases that the probability that the null hypothesis is true is very low. Simply put, if the null hypothesis is never true, there is no real point in testing it. O’Brien and Castelloe (2007) suggest an approach for using the results of null hypothesis tests to empirically estimate the possibility that the null hypothesis is true. In particular, they define what they call the “crucial” Type I and Type II error rates.

O’Brien and Castelloe (2007) turn traditional null hypothesis testing around to define their “crucial” rates. The crucial Type I error rate is the probability that the null hypothesis is true given that is has been rejected. A traditional Type I error involves rejecting the null when it is true. A crucial Type I error represents what is essentially the obverse of this traditional error – i.e., the likelihood that the null hypothesis is true given that it has been rejected. The crucial Type II error rate is the probability that the null hypothesis is false given that it has not been rejected.

Assume that you have not done a study yet. P(H0) represents the prior probability that the null hypothesis is true – i.e., the likelihood that it is true, independent of the results of any study. This prior probability is an important component of determining crucial Type I and crucial Type II errors. The probability of crucial Type I (a*) and crucial Type II (b*) errors are defined as:

$$a^* = [a\ (1-P(H0)]/[(\ a\ (1-P(H0))(1-b)]$$

$$b^* = [b\ (1-P(H0)]/[(\ b\ (1-P(H0))(1-a)(1-\ P(H0)]$$

As O’Brien and Castelloe (2007) note, when statistical power is high, both the crucial Type I error rate and the crucial Type II error rate are likely to be low.

The formulas above suggest that the definitions of crucial Type I and Type II errors revolve around the same issue that is critical in defining traditional Type I and Type II errors, the prior probability that the null hypothesis is true [P(H0)]. Unfortunately, this fact suggests that under certain circumstances, crucial Type I error calculations can be very misleading. Suppose, for example, that you know the traditional hypothesis is false (i.e., P(H0) = 0). If you are certain before you do a study that the null hypothesis is false, that does not change once you have done a study that rejects the null hypothesis. However, if P(H0) = 0, a* = [a]/[(a*(1-b)], which will almost certainly not be zero.

The formulas provided by O'Brien and Castelloe (2007) are most useful when there is some real chance that the null hypothesis is true, and you can estimate that probability with some precision. Unfortunately, this probably does not happen very often

End of Boxed Section

Summary

Tests of the traditional nil hypothesis, that treatments or interventions have no effect, have been criticized by methodologists. The most important criticism is that the nil hypothesis is, by definition, almost always wrong. As a result, tests of that hypothesis have little real value, and some of the procedures that are widely used in statistical analysis (e.g., adopting a stringent alpha level to control for Type I errors) are illogical and costly.

This chapter describes an alternative to the traditional nil hypothesis test, in which the null hypothesis describes a range of possible outcomes. In particular, it is often possible to define a range of effect sizes that would, by any reasonable standard, be defined as negligible. One possibility is to define any treatment effect that accounts for less than 1% of the variance in treatments as negligible (another possibility is to use 5% of the variance as a cutoff). Tests of the null hypothesis that treatment effects either fall within some range defined as negligible, or that the effects of treatments are large enough to be or real interest to the researcher are easy to perform and they offer numerous advantages. First, the results of these tests are not trivial. If a researcher can confidently reject the hypothesis that treatments had a negligibly small effect, he or she is left with the conclusion that they had effects large enough to be important.

The minimum-effects tests described here are simple extensions of the familiar *F* test. The only difference between nil hypothesis F tests and minimum-effect F tests is that they use different *F* tables. Nil hypothesis tests are based on the hypothesis that treatments have no effect, a hypothesis that is extremely unlikely to be true. Minimum-effect tests have the advantage of being simple and informative. In all of the sections that follow, we will present minimum-effect tests as well as nil hypothesis tests. In our final chapter, we will argue that the advantages of alternatives to nil hypothesis tests, such as the minimum-effects test, are so compelling that they call into question the continuing use of the traditional nil hypothesis test.

Footnotes

1. Remember, the equation for going from *F* to *PV,* which is presented in Chapter 2, is
$PV = (\mathrm{df}_{hyp} * F) / [(\mathrm{df}_{hyp} * F) + \mathrm{df}_{err}]$, which in this case equals
$PV = (3 * 2.46) / [(3 * 2.46) + 50] = .128$, which rounds to 13% of the variance explained